

Introduction to Machine Learning for the Life Sciences

John Joseph Valletta

University of Exeter, Penryn Campus, UK

1st-2nd June 2015

www.exeter.ac.uk/as/rdp/

Housekeeping

- To log onto computers on the Penryn campus you need to add the domain **isad**, for e.g isad\jjv207
- For Wi-Fi log on **UoE_Secure** if you have University of Exeter credentials or **FXNET** otherwise
- **All** catering will be served in the **GREEN** seminar room situated one floor up from the Exchange Lecture Theatre
- The IT Training Suite (lab sessions) is accessed from the main library entrance (same floor as the Lecture Theatre). Follow signs to IT Room (a right, then another right)
- The University would appreciate greatly if you do not bring food or drinks in the Exchange Lecture Theatre and IT Training Suite
- If you get lost ask the locals or poke me on 07847222721

www.exeter.ac.uk/as/rdp/

Housekeeping

Day 1	1st June 2015
10:00 - 10:30	Arrival + Tea + Coffee
10:30 - 11:30	<i>Talk:</i> Introduction to Machine Learning
11:30 - 12:30	<i>Talk:</i> Clustering (unsupervised learning)
12:30 - 13:30	Lunch
13:30 - 15:00	Lab Session #1 (E/DM Exeter IT Training Suite)
15:00 - 15:30	Tea + Coffee
15:30 - 16:00	<i>Contributed Talk</i> Finlay Maguire: Stumbling Over the Decision Boundary
16:00 - 16:30	<i>Contributed Talk</i> Colin Torney: Identification of Wildebeest from Aerial Photos
16:30 - 17:00	Recap of Day 1

www.exeter.ac.uk/as/rdp/

Housekeeping

Day 2	2nd June 2015
09:00 - 09:30	Tea + Coffee
09:30 - 10:30	<i>Talk:</i> Classification (supervised learning)
10:30 - 11:00	<i>Contributed Talk</i> Finlay Maguire: Casting a Deep Net: Classifying Plankton from Images
11:00 - 11:30	Tea + Coffee
11:30 - 13:00	Lab Session #2 (E/DM Exeter IT Training Suite)
13:00 - 14:00	Lunch
14:00 - 15:00	<i>Contributed Talk</i> Richard Everson: Cows, Crickets and Bees: Learning from Video
15:00 - 16:00	<i>Talk:</i> Practical Issues and Final Discussion

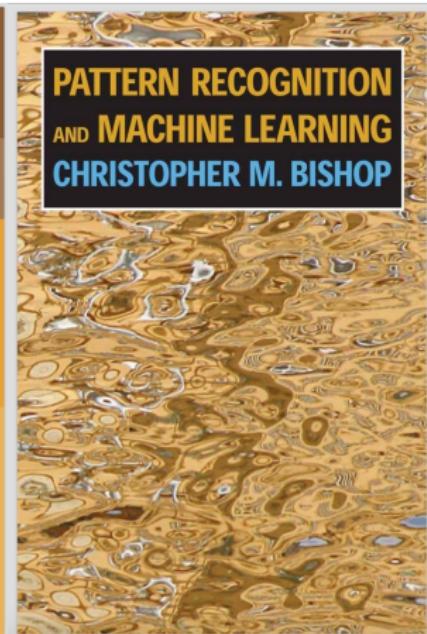
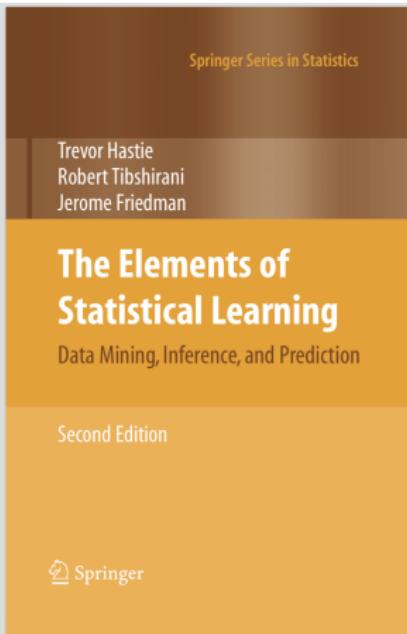
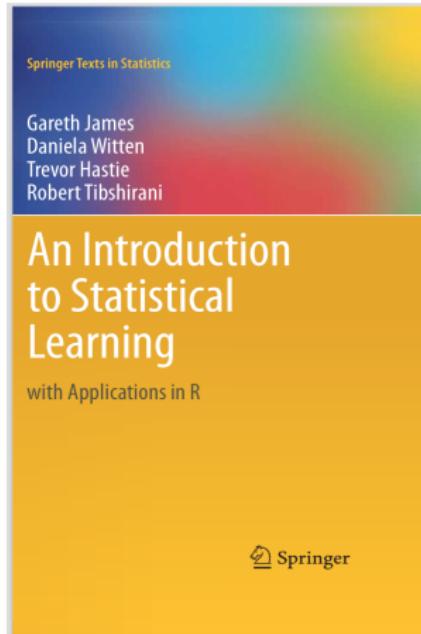
www.exeter.ac.uk/as/rdp/

Workshop learning outcomes

- Understand the key concepts and terminology used in the field of machine learning
- Build predictive models for clustering and classification problems
- Apply machine learning algorithms in R to a variety of real-world datasets
- Recognise practical issues in data-driven modelling

www.exeter.ac.uk/as/rdp/

Bibliography



www.exeter.ac.uk/as/rdp/

Overview

- Why are we here?
- What is machine learning?
- Types of machine learning methods
- Statistics vs Machine Learning
- Terminology
- A bird's-eye view of machine learning

Overview

- Why are we here?
- What is machine learning?
- Types of machine learning methods
- Statistics vs Machine Learning
- Terminology
- A bird's-eye view of machine learning

Overview

- Why are we here?
- What is machine learning?
- Types of machine learning methods
- Statistics vs Machine Learning
- Terminology
- A bird's-eye view of machine learning

Overview

- Why are we here?
- What is machine learning?
- Types of machine learning methods
- Statistics vs Machine Learning
- Terminology
- A bird's-eye view of machine learning

Overview

- Why are we here?
- What is machine learning?
- Types of machine learning methods
- Statistics vs Machine Learning
- Terminology
- A bird's-eye view of machine learning

Overview

- Why are we here?
- What is machine learning?
- Types of machine learning methods
- Statistics vs Machine Learning
- Terminology
- A bird's-eye view of machine learning

Overview

- Why are we here?
- What is machine learning?
- Types of machine learning methods
- Statistics vs Machine Learning
- Terminology
- A bird's-eye view of machine learning

Why are we here?

The infamous Big Data!

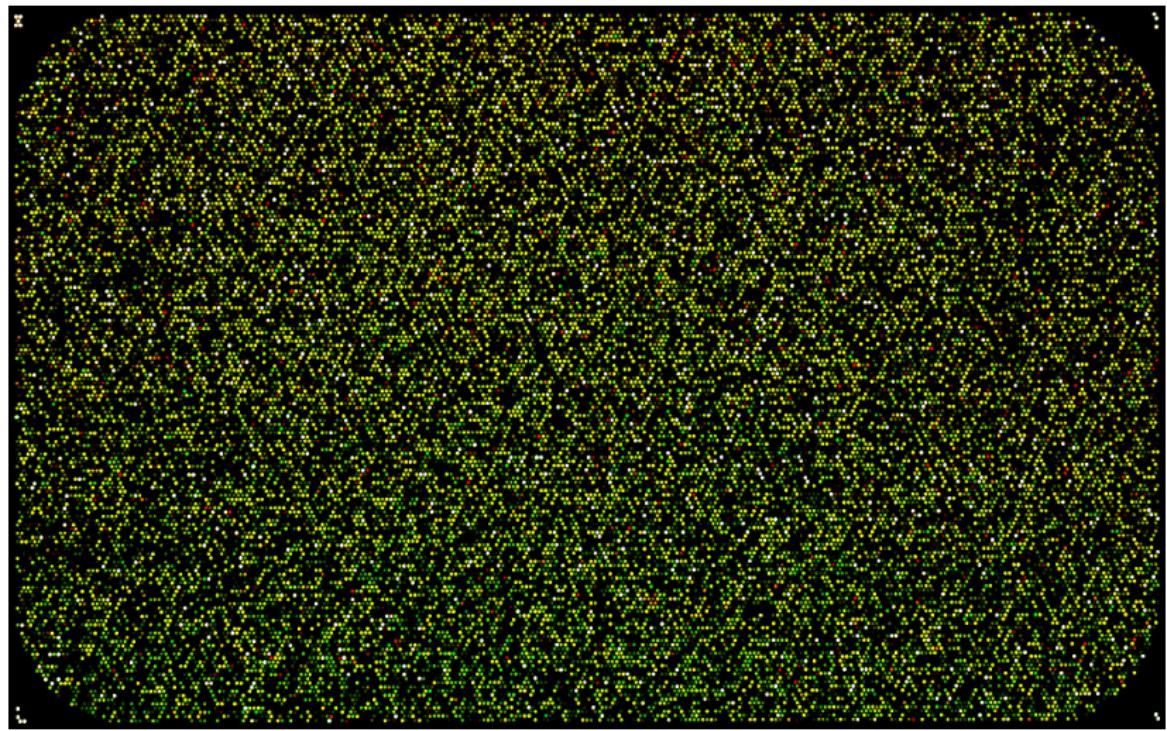
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	
	BrdIndx	Area	Round	Bright	Compact	ShipIndx	Mean_G	Mean_R	Mean_NIR	SD_G	SD_R	SD_NIR	LW	GLCM1	Rect	GLCM2	Dens	Assym	NDVI	BordLength	GLCM3	
1	class																					
2	car	1.27	91	0.97	231.38	1.39	1.47	207.92	241.74	244.85	21.41	20.4	18.69	2.19	0.48	0.87	6.23	1.6	0.74	-0.08	56	
3	concrete	2.36	241	1.56	216.15	2.46	2.51	187.85	229.39	231.2	6.57	6.97	7.02	1.52	0.45	0.63	6.35	1.55	0.69	-0.1	156	
4	concrete	2.12	266	1.47	232.18	2.07	2.21	206.54	242.42	245.79	6.16	4.93	5.53	1.14	0.52	0.78	6.19	1.88	0.19	-0.08	144	
5	concrete	2.42	399	1.28	230.4	2.49	2.73	204.6	243.27	243.32	5.76	5.56	5.46	2.51	0.5	0.74	6.28	1.51	0.83	-0.09	218	
6	concrete	2.15	944	1.73	193.18	2.28	4.1	165.98	205.55	208	11.46	8.9	9.77	12.26	0.71	0.59	7.36	0.63	0.99	-0.11	504	
7	tree	3.11	169	1.47	172.22	2.49	3.35	240.18	127.65	148.83	8.41	10.34	11.5	1.87	0.43	0.55	6.44	1.38	0.7	0.31	174	
8	car	1.2	44	0.79	208.8	1.14	1.36	180.95	221.61	221.82	35.42	36.45	35.17	2	0.54	0.89	5.84	1.58	0.63	-0.1	36	
9	car	1	88	0.22	234.51	1.11	1.12	208.5	246.48	246.55	12.3	11.29	8.81	2	0.44	0.93	6.08	1.78	0.5	-0.08	42	
10	building	1.59	1737	0.67	231.61	1.3	1.64	186.86	230.64	239.13	7.08	7.03	7.28	1.49	0.47	0.9	6.51	2.15	0.51	-0.11	274	
11	tree	2.37	153	1.3	120.24	2.85	2.59	184.45	181.5	95.06	10.21	11.99	11.3	2.05	0.56	0.64	6.62	1.57	0.5	0.39	128	
12	building	1.57	3552	0.46	212.23	1.24	1.24	173.03	225.94	225.8	5.4	5.34	5.63	1.4	0.11	0.4	6.24	2.3	0.31	-0.14	384	
13	asphalt	4.19	418	0.48	83.35	4.21	4.3	68.07	88.8	91.17	6.61	6.3	6.12	1.1	0.46	0.44	5.52	1.31	0.36	-0.13	352	
14	building	1.3	104	0.58	183.93	1.39	1.64	169.57	206.61	172.6	6.15	6.37	6.8	3.95	0.38	0.89	6.46	1.63	0.86	-0.11	210	
15	grass	1.14	289	0.38	173.16	1.21	1.21	213.73	145.56	160.23	10.3	11.55	11.1	1.79	0.54	0.91	6.77	2.02	0.52	0.39	82	
16	shadow	2.03	249	1	39	60.2	2.2	21.36	35.86	38.92	44.07	7.72	6.36	6.87	1.35	0.63	0.76	6.42	1.92	0.3	-0.04	134
17	building	1.31	1639	0.68	205.31	1.26	1.37	196.24	230.05	180.65	7.15	7.82	8.69	0.49	0.9	0.53	6.53	2.07	0.56	-0.1	222	
18	tree	2.68	285	1.48	138.01	2.53	2.7	201.84	98.06	114.13	15.24	15.1	14.31	1.01	0.7	0.63	7.28	1.75	0.15	0.35	182	
19	soil	2.79	293	1.37	237.11	3.4	3.1	227.22	243.45	241.68	7.9	9	9.86	2.17	0.45	0.67	6.66	1.75	0.73	-0.03	212	
20	building	1.21	2707	0.78	244.7	1.34	1.12	229.52	252.11	252.37	7.5	4.98	4.42	1.24	0.48	0.85	6.64	2.2	0.4	-0.05	260	
21	shadow	1.13	237	0.32	41.25	1.22	1.22	47.71	34.69	41.55	8.79	8.29	7.77	2.18	0.48	0.92	6.41	1.84	0.68	0.16	72	
22	pool	1.28	173	0.85	157.34	1.53	1.41	85.43	161.11	225.49	14.8	12.18	16.76	2.18	0.77	0.76	6.64	1.8	0.65	-0.31	74	
23	shadow	3.23	203	2.15	66.75	3.55	3.4	60.3	68.47	71.47	14.52	16.39	14.8	1.46	0.52	0.31	6.83	1.25	0.47	-0.06	194	
24	concrete	2.09	178	1.63	215.76	2.4	2.66	188.69	229.03	229.55	5.25	4.95	5.01	3.55	0.54	0.59	6.07	1.06	0.94	-0.1	142	
25	tree	2	138	1.31	154.97	2.46	2.04	200.16	144.49	140.25	13.52	13	12.97	1.18	0.58	0.73	6.74	1.7	0.45	0.23	96	
26	grass	1.79	203	1.83	152.09	2.41	2.46	193.34	232.78	139.15	14.03	13.18	10.17	4.9	0.56	0.67	6.91	0.96	0.92	0.22	140	
27	concrete	2	792	1	238.34	1.88	2.24	211.08	251.96	251.97	6.43	3.73	3.78	2.49	0.51	0.84	6.4	1.69	0.81	-0.09	252	
28	grass	3.54	429	1.79	176.66	2.73	3.93	205.01	160.12	164.84	7.52	7.29	7.02	2.28	0.57	0.58	6.58	1.33	0.8	0.12	326	
29	building	1.38	1078	0.64	243.74	1.27	1.43	226.39	252.62	252.21	7.17	7.38	4.43	1.42	0.48	0.9	6.58	2.17	0.37	-0.05	188	
30	building	1.92	956	0.75	173.53	1.76	2.05	161.19	201.82	162.26	7.79	7.24	6.78	2	0.51	0.8	6.67	1.92	0.62	-0.11	254	
31	building	2.74	316	1.92	200.53	2.73	3.32	181.09	208.44	212.05	7.5	8.4	8.22	3.38	0.5	0.57	6.49	1.04	0.93	-0.07	236	
32	asphalt	2.94	642	1.43	84.16	2.56	3.02	67.13	90.65	94.7	7.06	7.08	7.15	1.34	0.55	0.66	6.56	1.82	0.33	-0.15	306	
33	shadow	1.39	212	0.86	49.89	1.43	1.79	61.68	40.14	47.87	9.51	7.03	7.09	3.89	0.53	0.84	6.54	1.35	0.92	0.21	106	
34	building	1.48	3084	0.93	230.71	1.33	1.33	215.62	252.62	252.84	7.04	3.65	7.07	9.33	0.53	0.89	6.65	1	0.98	-0.08	560	
35	grass	2.44	554	1.72	146.12	3.03	2.49	213.73	102.01	122.61	8.67	6.99	7.24	1.05	0.65	0.65	6.7	1.72	0.37	0.35	234	
36	tree	2.89	288	1.41	124.98	2.4	2.98	187.27	88.61	99.07	15.77	13.45	13.66	1.3	0.69	0.62	7.25	1.76	0.12	0.36	202	
37	building	1.4	1278	0.73	226.77	1.42	2.06	204.75	251.55	234.01	7.03	4.58	7.13	6.29	0.46	0.91	6.61	1.22	0.96	-0.1	254	
38	asphalt	2.66	998	1.34	71.58	2.35	2.82	57.98	75.54	81.23	8.51	8.88	8.99	1.82	0.49	0.73	6.88	1.77	0.66	-0.13	356	
39	tree	1.66	328	0.9	237.5	1.73	1.74	211.47	250.57	250.48	6.5	4.95	4.99	1.6	0.47	0.83	6.38	1.94	0.54	-0.08	126	
40	tree	3.2	373	1.61	148.7	2.78	3.73	213.41	104.99	125.76	13.31	10.07	10.96	2.58	0.68	0.57	7.33	1.3	0.81	0.33	288	
41	concrete	1.41	1046	0.76	239.94	1.37	1.9	215.11	254.27	252.25	6.2	3.72	3.71	4.94	0.36	0.9	6.4	1.39	0.93	-0.08	246	
42	soil	2.26	397	1.63	236.16	3.43	2.38	233.62	238.42	236.43	8.41	10.71	11.52	1.62	0.47	0.69	6.75	1.66	0.45	-0.01	190	
43	tree	1.96	144	0.98	112.87	2.31	2.13	168.78	78.85	90.97	11.45	9.01	8.27	1.88	0.69	0.72	6.54	1.65	0.58	0.36	102	
44	building	1.51	869	1.17	189.67	1.56	1.71	164.41	208.08	196.5	7.45	8.08	7.53	2.57	0.54	0.78	6.66	1.57	0.84	-0.12	202	
45	building	1.87	3645	0.75	236.75	1.55	1.97	204.59	252.82	252.83	5.72	3.22	3.29	1.78	0.24	0.86	6.32	2.05	0.62	-0.11	476	
46	building	1.55	2306	0.74	211.82	1.29	1.71	185.81	237.23	216.93	6.91	6.28	6.21	2.43	0.42	0.9	6.55	2	0.62	-0.11	328	
47	grass	1.83	183	0.83	144.61	1.97	1.96	156.36	139.63	137.83	7.98	7.32	7.61	1.6	0.46	0.78	6.52	1.75	0.62	0.06	106	
48	grass	3.72	557	2.13	152.42	3.58	3.94	166.87	145.44	144.96	7.65	7.94	7.7	1.79	0.58	0.55	6.74	1.39	0.52	0.07	372	
49	shadow	1.72	663	0.43	45.19	1.7	1.81	47.41	41.36	47.11	9.06	7.43	7.24	1.55	0.7	0.87	6.72	2.13	0.43	0.06	186	
50	soil	2.23	224	1.23	239.87	2.24	2.31	225.22	247.39	247	6.67	6.04	6.51	1.41	0.42	0.63	6.3	1.64	0.57	-0.05	138	

Normal View Ready Sum=0

No. of observations << No. of predictors ($n \ll p$ problem)

Why are we here?

The infamous **Big Data!**



No. of observations \ll No. of predictors ($n \ll p$ problem)

Water, water everywhere nor any drop to drink

LET'S SOLVE THIS PROBLEM BY
USING THE BIG DATA NONE
OF US HAVE THE SLIGHTEST
IDEA WHAT TO DO WITH



© marketoonist.com

Water, water everywhere nor any drop to drink

SAY BIG DATA

ONE MORE TIME

memegenerator.net

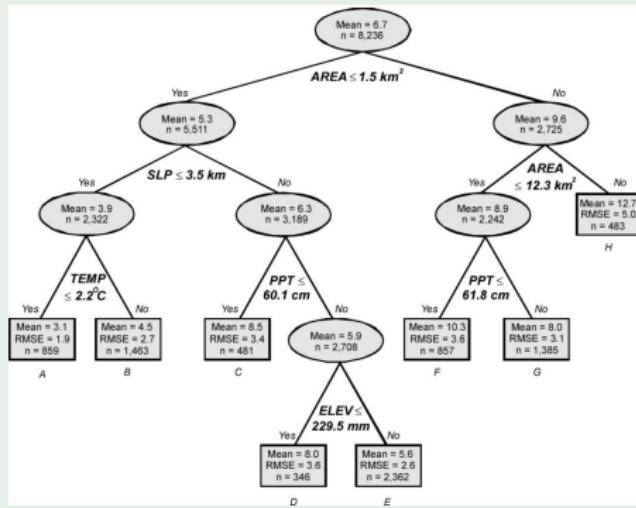
What's the problem?

Predicting fish species richness

Olden *et al.* Q Rev Biol 2008, 83(2):171-193

Data: Lake surface area, shoreline perimeter, air temperature, precipitation and elevation

Method: Decision trees (supervised)



What's the problem?

Detection of malarial parasites

Purwar *et al.* Malar J 2011, 10:364

Data: Image intensity

Method: Modified k-means clustering (unsupervised)

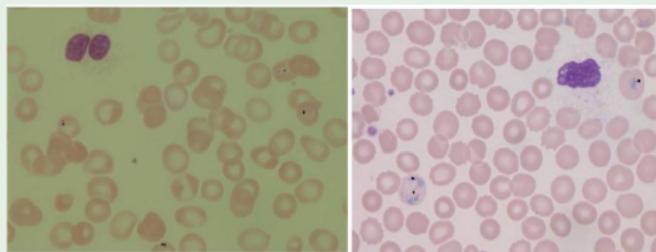
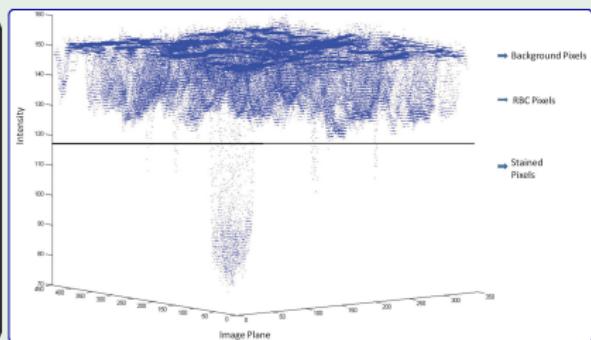


Figure 17 Parasites marked image.



What's the problem?

Creating carbon-density maps

Baccini *et al.* Nature Clim. Change 2012, 2:182-185

Data: Light detection and ranging (LiDAR) (elevation data)

Method: Random forests (ensemble of decision trees) (supervised)

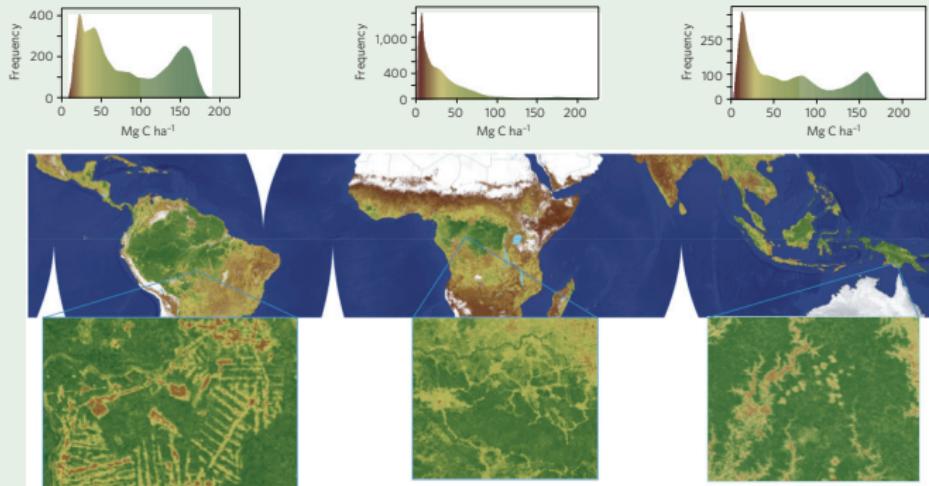


Figure 1 | Carbon contained in the aboveground live woody vegetation of tropical America, Africa and Asia (Australia excluded). The upper panels show the frequency distribution of carbon in units of Mg C ha^{-1} for each region. Inset figures across the bottom provide higher-resolution examples of the spatial detail present in the satellite-derived biomass data set. Carbon amount is represented in the maps as a colour scheme from dark brown (low carbon) to dark green (high carbon). See upper panels for numeric values.

What's the problem?

Acoustic classification of multiple simultaneous bird species

Briggs et al. J Acoust Soc Am 2012, 131(6):4640-4650

Data: Segments in spectrogram (time vs frequency) from 10 secs audio recordings (corresponding to syllables of bird call)

Method: Multi-instance multi-label learning (supervised)

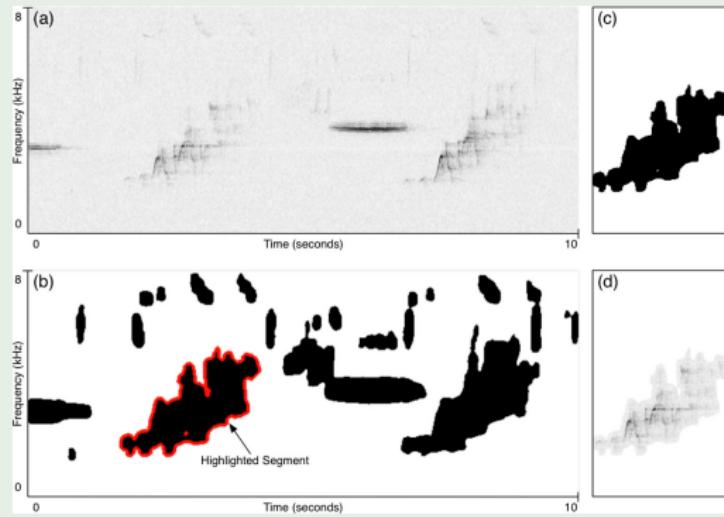


FIG. 3. (Color online) Extracting a syllable from the segmentation results. (a) The original spectrogram, (b) the binary mask generated by our segmentation algorithm. The highlighted segment will be further processed in this example. Note that several other segments overlap in time. (c) A cropped mask of the highlighted segment. (d) The masked and cropped spectrogram corresponding to the highlighted segment.

What my mum thinks machine learning is



**Artificial
Intelligence**

Terminator - Rise of The Machines

Who uses machine learning?

Google NETFLIX

SAY HELLO TO
KINECT
FOR XBOX 360



You Tube

amazon.com[®]

Who uses machine learning?

Machine Learning in Ecosystem Informatics and Sustainability

Thomas G. Dietterich

School of Electrical Engineering and Computer Science
Oregon State University
tgd@cs.orst.edu

MACHINE LEARNING IN THE LIFE SCIENCES



©BRAND X, PHOTODISC

Machine Learning in the Life Sciences

How it is Used on a Wide Variety of Medical Problems and Data

KRZYSZTOF J. CIOS, LUKASZ A. KURGAN,
AND MAREK REFORMAT

VOLUME 83, No. 2

THE QUARTERLY REVIEW OF BIOLOGY

JUNE 2008



MACHINE LEARNING METHODS WITHOUT TEARS: A PRIMER
FOR ECOLOGISTS

Data Analysis and Mining in the Life Sciences

Nam Huyn

SurroMed, Inc.

2375 Garcia Ave, Mountain View, CA 94043, USA
phuyn@surreomed.com

There are even competitions now!



[Sign Up](#) [In the News](#) [Judging Panel](#) [Visit HPN](#)

0 0 1 0 1 1 0 1 0 1 1 0 1 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 1 1 0 0 0

[Dashboard](#)

[Home](#)

[Data](#)

[Information](#)

- Description
- Evaluation
- Rules
- Dos and Don'ts
- FAQ
- Milestone Winners
- Timeline

[Forum](#)

[Leaderboard](#)

- Public
- Private

[Leaderboard](#)

1. POWERDOT
2. EXL Analytics

Improve Healthcare, Win \$3,000,000.

Identify patients who will be admitted to a hospital within the next year using historical claims data. (Enter by 06:59:59 UTC Oct 4 2012)

Please note: This competition is over! The leaderboard now displays the final results.

Lots of them actually...

Active Competitions			
		March Machine Learning Mania 2015 Predict the 2015 NCAA Basketball Tournament	29 days 108 teams \$15,000
		National Data Science Bowl Predict ocean health, one plankton at a time	31 days 661 teams \$175,000
		Driver Telematics Analysis Use telematic data to identify a driver signature	31 days 1024 teams \$30,000
		BCI Challenge @ NER 2015 A spell on you if you cannot detect errors!	11 days 238 teams \$1,000
		Microsoft Malware Classification Challenge (B...) Classify malware into families based on file content and characteristics	2 months 77 teams \$16,000
		How much did it rain? Predict probabilistic distribution of hourly rain given polarimetric radar measurements	3 months 79 teams \$500
		Sentiment Analysis on Movie Reviews Classify the sentiment of sentences from the Rotten Tomatoes dataset	15 days 805 teams Knowledge

Source: <http://www.kaggle.com/>

So what is machine learning?

So what is machine learning?

A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E

— Tom Mitchell

So what is machine learning?

A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E

— Tom Mitchell

A scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model from example inputs and using that to make predictions or decisions, rather than following strictly static program instructions.

— Wikipedia

So what is machine learning?

A machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably improves its performance P at task T, following experience E

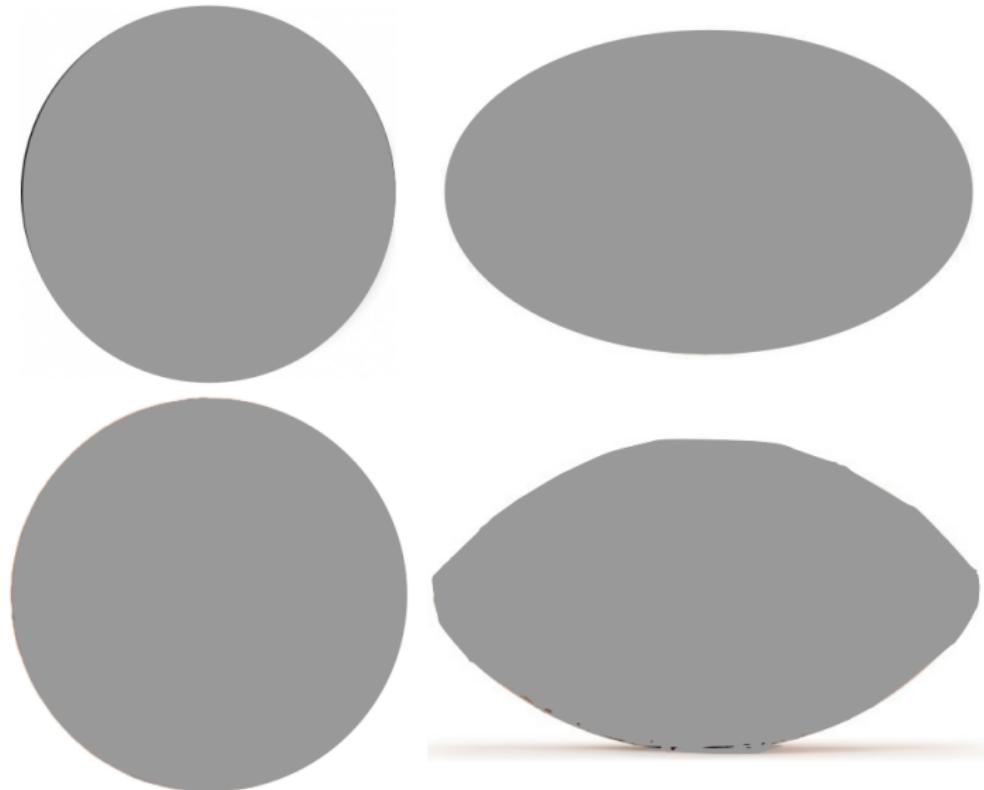
— Tom Mitchell

A scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model from example inputs and using that to make predictions or decisions, rather than following strictly static program instructions.

— Wikipedia

Machines learn using flashcards

Group by shape (unsupervised learning)



Add labels (supervised learning)



Types of machine learning methods: Unsupervised learning

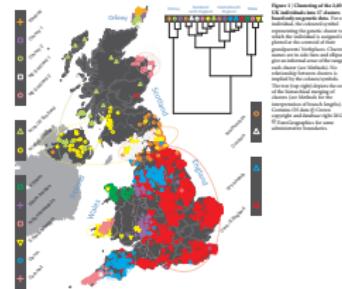
Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes
- **Density Estimation** - determine the distribution of data
- **Dimensionality Reduction** - identify and remove redundant dimensions

Types of machine learning methods: Unsupervised learning

Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes



- **Density Estimation** - determine the distribution of data
- **Dimensionality Reduction** - identify and remove redundant dimensions

Types of machine learning methods: Unsupervised learning

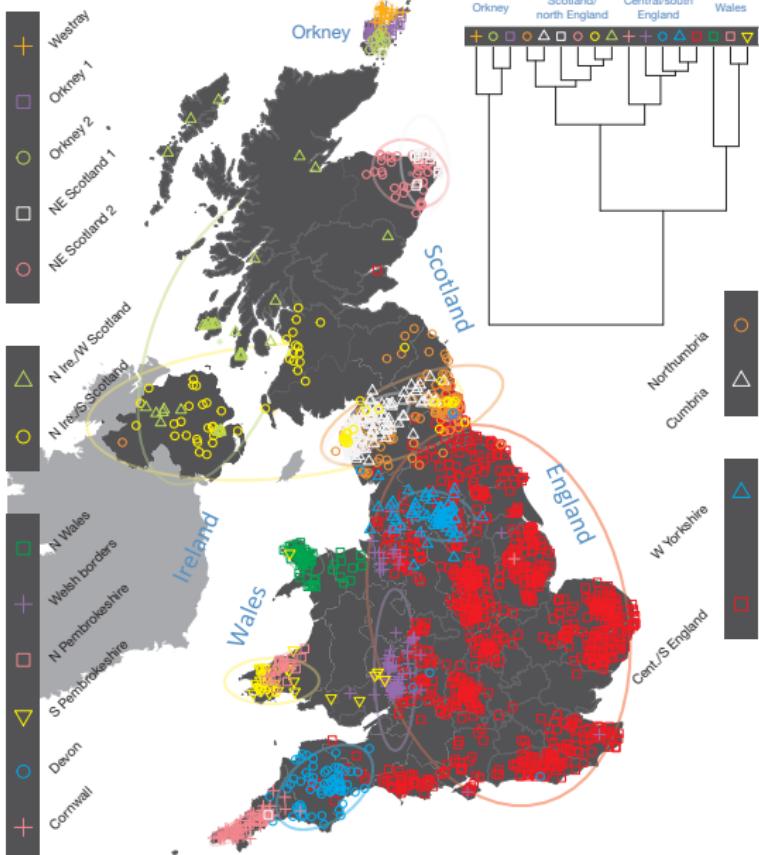
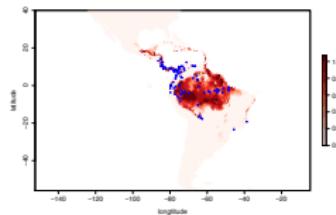
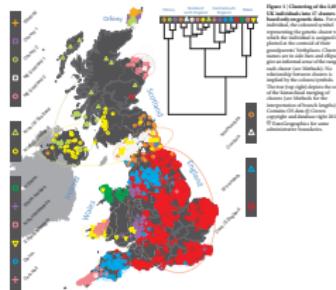


Figure 1 | Clustering of the 2,039 UK individuals into 17 clusters based only on genetic data. For each individual, the coloured symbol representing the genetic cluster to which the individual is assigned is plotted at the centroid of their grandparents' birthplaces. Cluster names are in side-bars and ellipses give an informal sense of the range of each cluster (see Methods). No relationship between clusters is implied by the colours/symbols. The tree (top right) depicts the order of the hierarchical merging of clusters (see Methods for the interpretation of branch lengths). Contains OS data © Crown copyright and database right 2012. © EuroGeographics for some administrative boundaries.

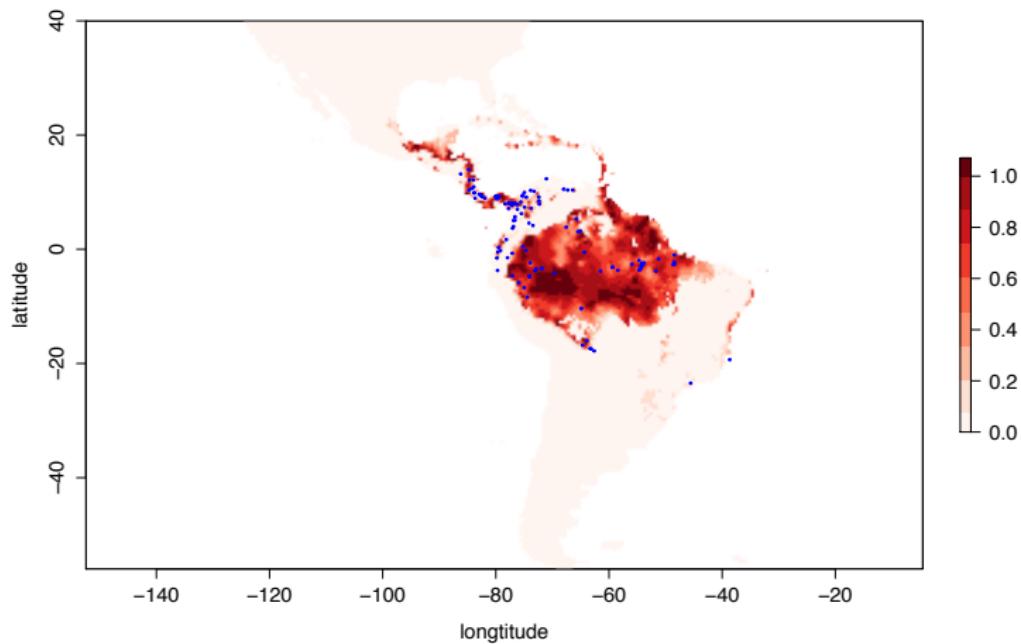
Types of machine learning methods: Unsupervised learning

Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes
 - **Density Estimation** - determine the distribution of data
 - **Dimensionality Reduction** - identify and remove redundant dimensions



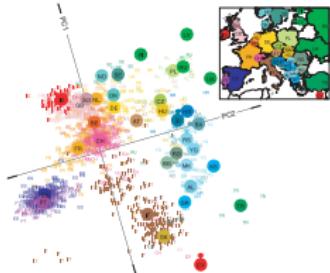
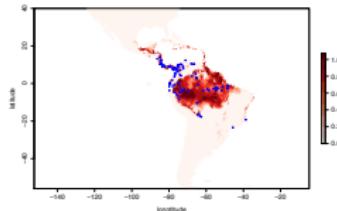
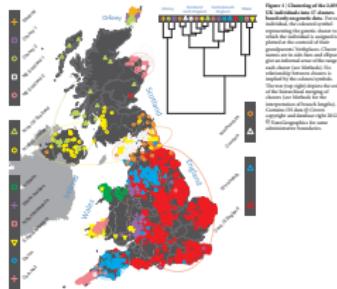
Types of machine learning methods: Unsupervised learning



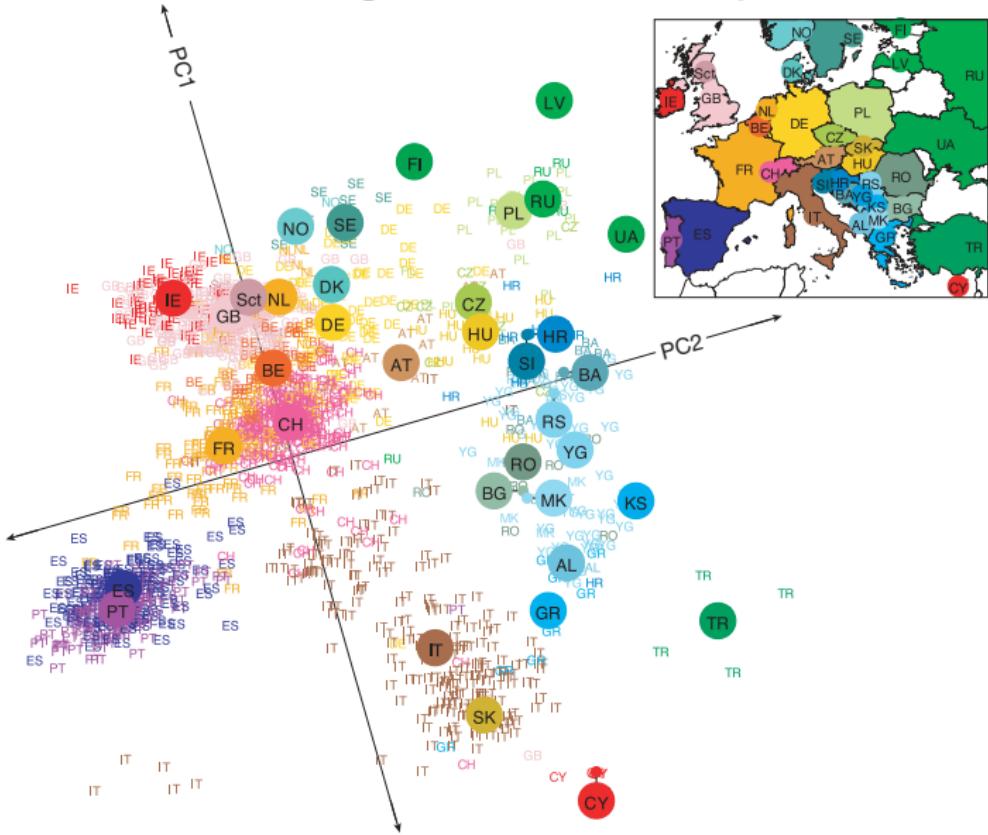
Types of machine learning methods: Unsupervised learning

Inputs have *no* corresponding output labels

- **Clustering** - discovering groups having similar attributes
- **Density Estimation** - determine the distribution of data
- **Dimensionality Reduction** - identify and remove redundant dimensions



Types of machine learning methods: Unsupervised learning



Types of machine learning methods: Supervised learning

Inputs have corresponding output labels

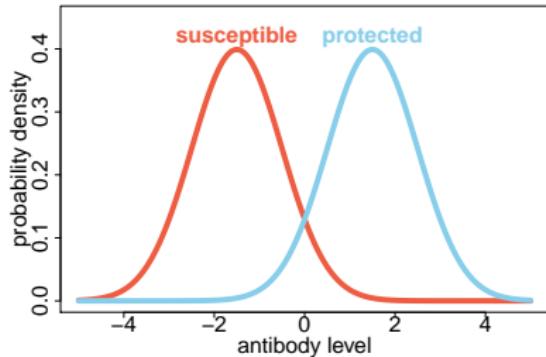
- Classification - output is categorical
- Regression - output is continuous

Types of machine learning methods: Supervised learning

Inputs have corresponding output labels

- **Classification** - output is categorical

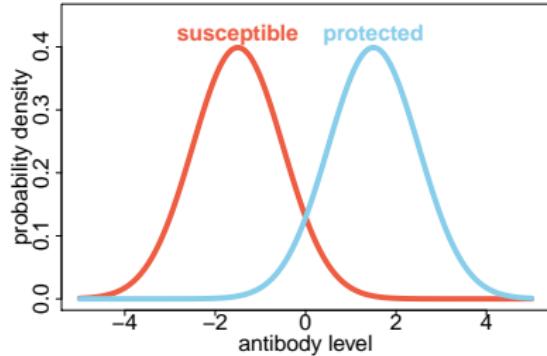
- **Regression** - output is continuous



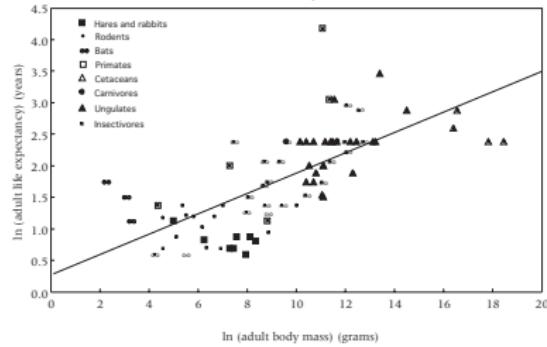
Types of machine learning methods: Supervised learning

Inputs have corresponding output labels

- **Classification** - output is categorical



- **Regression** - output is continuous

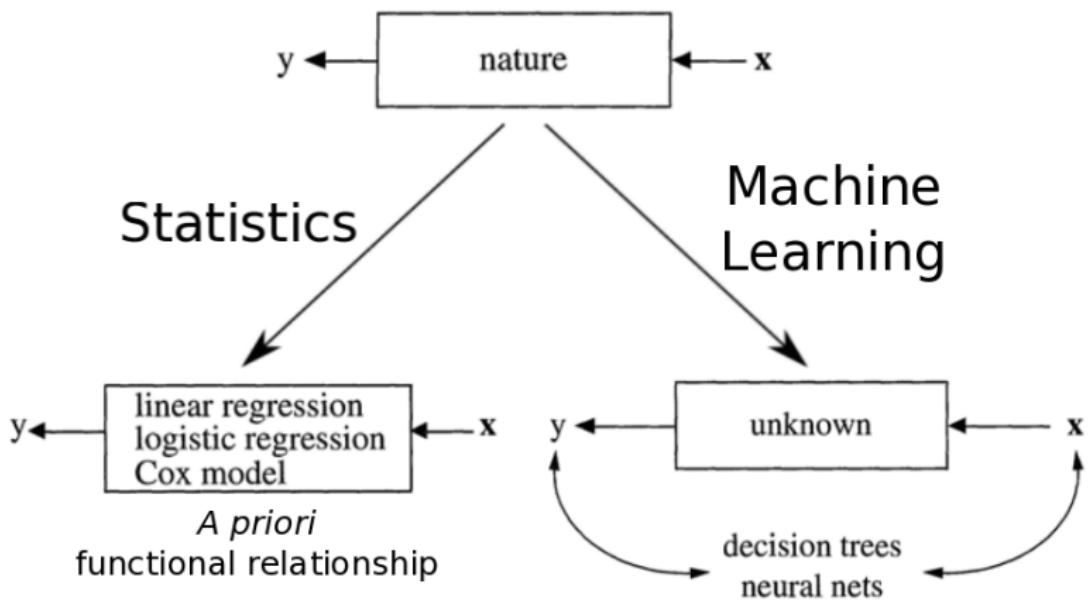


Statistics vs Machine Learning (not mutually exclusive)

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman



Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** – provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2 , chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2 , Cp statistic, AIC)

Machine Learning

- **Philosophy** – replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** – how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2 , chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2 , Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2 , chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2 , Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2 , chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2 , Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2 , chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2 , Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2 , chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2 , Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics vs Machine Learning (not mutually exclusive)

Statistics

- **Philosophy** - provide humans a set of data analysis tools
- **Focus** - what is the relationship between the data and the outcome?
- **Inference** - how was the observed data generated
- **Learning** - All measured data then perform inference on the population
- **Validation** - Measures of fit (R^2 , chi-square test)
- **Selection** - Adjusted measures of fit (adjusted R^2 , Cp statistic, AIC)

Machine Learning

- **Philosophy** - replace humans in the processing of data
- **Focus** - how can we predict the outcome using the data?
- **Prediction** - how can we use observed data to predict the future
- **Learning** - Training dataset then perform predictions on testing dataset
- **Validation** - How well it predicts "unseen" data (generalisation)
- **Selection** - Cross-validation and out-of-bag errors

Statistics and machine learning complement each other

The best solution could be an algorithmic model (machine learning), or maybe a data model, or maybe a combination. But the trick to being a **scientist** is to be open to using a wide variety of tools.

— Leo Breiman

Machine learning extends a scientist's toolbox to deal with the explosion of data that the life sciences have seen in recent years due to the changing nature of science and technology

The objective is not just to get a better fit to the data but to have a predictive model that *generalises* well, that is, gives good predictions to *unseen* data

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe "well" our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

Terminology

Training Dataset: Used to train a set of models

Validation Dataset: Used for model selection and validation. Helps us to select a parsimonious model i.e a model which is complex enough to describe “well” our data but not more complex

Testing Dataset: Used to compute the *generalisation* error. Evaluate model performance on previously unseen data

Inputs: Covariates, Predictors, Features, Attributes

Training error: In sample error, Resubstitution error

Testing error: Out of sample error, Generalisation error

A bird's-eye view of machine learning

